Guo *et al. J Infrastruct Preserv Resil* (2023) 4:4
https://doi.org/10.1186/s43065-023-00071-y

Journal of Infrastructure Preservation and Resilience

RESEARCH

Open Access

Enclosing contour tracking of highway construction equipment based on orientation-aware bounding box using UAV

Yapeng Guo[1], Yang Xu[2], Zhonglong Li[1], Hui Li[2] and Shunlong Li[1*]

Abstract

Construction equipment tracking of highway construction site can obtain the spatiotemporal location in real time and provide data basis for construction risk control. The complete 2D moving of construction equipment in surveillance videos could be spatially represented by the translation, rotation and size change of corresponding images. To describe the temporal relationships of these variables, this study proposes a construction equipment enclosing contour tracking method based on orientation-aware bounding box (OABB), where UAV surveillance videos are employed to alleviate the occlusion problem. The method balances the rotation insensitivity of horizontal bounding box and the complexity of pixel-level segmented contour, which has three modules. The first module integrates OABB into a deep learning detector to provide detected contours. The second module updates OABBs with Kalman prediction to output tracked contours. The third module manages IDs of multiple tracked contours for construction equipment motions. Five in-situ UAV videos including 4325 frames were employed as the evaluation dataset. The tracking performance achieved 2.657 degrees in angle error, 97.523% in MOTA and 83.243% in MOTP.

Keywords Construction equipment tracking, UAV surveillance videos, Highway construction site, Orientation-aware bounding box, Rotating angle

Introduction

The motion of construction equipment in the 2D plane based on computer vision can be defined by translation and rotation. Considering that the distance from the photography plane to construction equipment might change, pixel size of corresponding equipment image also needs to be included. These constitute a complete 2D spatial description of the plane moving pattern of construction equipment, which is represented by the enclosing contour in this study. Precise spatial–temporal information of construction equipment is one of the most important datatypes in construction sites [1–3], which can be used to provide location feedback for equipment engaged in hazardous operations and early warning for construction personnel around the equipment. Furthermore, such information can provide the basis for the organization and guidance of traffic flow at key nodes of construction sites and for the analysis of working productivity efficiency [4, 5]. Enclosing contour tracking of construction equipment, used for relatively precise spatial–temporal information acquisition, has become critical to improve efficiency and ensure safety in construction sites.

Kinematic-based construction equipment tracking methods using installed devices (e.g., radio frequency identification, global positioning systems, ultra-wideband, Bluetooth low-energy, accelerator) [2–12] have been validated with good accuracy and real-time processing speed for moving trajectories extraction. In addition to those approaches, vision-based

*Correspondence:
Shunlong Li
lishunlong@hit.edu.cn
[1] School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China
[2] School of Civil Engineering, Harbin Institute of Technology, Harbin 150090, China

Springer Open

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

sensing methods have become promising due to non-contact, low cost and abundant data. Many methods have been conducted treating equipment as a point, i.e., trajectory identification), including 2D trajectory [13–16] and 3D trajectories [17, 18]. These methods concentrate on the translations of construction equipment, but when the construction equipment is close to each other or close to the workers, its volume cannot be ignored. Therefore, the identification of more accurate information of construction equipment has attracted the attention of researchers, i.e., treating equipment as an enclosing contour.

Using horizontal bounding box (HBB) to represent the construction equipment enclosing contour and track the size (width and height) in addition to the translation (centre point coordinates) can alleviate the above limitations. HBB-based construction equipment enclosing contour tracking methods can detect rough equipment regions [19–24]. However, HBB has no rotation sensitivity, and its region contains a large number of non-equipment parts. Pixel-level segmented contour tracking is an appropriate way to accurately represent the construction equipment spatial–temporal information [1]. But robust segmented contour tracking based on deep learning needs complex manual labelling and temporal contour association, which would be superfluous for the 2D spatio-temporal description.

Thus, to balance the rotation insensitivity of the HBB and the high calculation complexity of pixel-level segmented contour, this study proposes an enclosing contour tracking method for construction equipment based on OABB using UAV surveillance videos. This study is arranged as follows: Sect. " Literature review" presents a literature review on vision-based tracking for construction equipment and arbitrary-oriented object detection; Sect. " Methodology" illustrates the methodology of the proposed approach; Sect. " Evaluation and implementation details" describes the dataset used to evaluate the algorithm, the evaluation metrics and the implementation details; Sect. " Results and discussions" shows the tracking results both qualitatively and quantitatively, with a discussion of the key update factor; Sect. " Conclusions and future works" concludes the research.

Literature review

In this section, tracking methods on vision-based for construction equipment will be reviewed. Because this research integrates OABB into the tracking method, research work in the field of arbitrary-oriented object detection will also be reviewed comprehensively.

Vision-based tracking methods for construction equipment

Many studies on construction equipment tracking based on computer vision techniques have been conducted. Some of them focus on the translation (moving trajectory) identification which treat the construction equipment as one point. Kim et al. [13] presented a mobile construction equipment 2D trajectory extraction method based on deep learning detector and image rectification technique using UAV videos. Tang et al. [14] took 2D tracks of construction equipment and predicted their locations using long short-term memory network and mixture density network. Zhao et al. [15] proposed a construction equipment tracking for 2D trajectory extraction using deep learning. Zhu et al. [16] proposed a particle filter-based construction equipment tracking method to acquire 2D trajectories. To calculate more accurate spatial locations of construction equipment, they [18] also developed a novel Kalman filter-based tracking method to estimate 3D positions using stereo vision. Jog. et al. [17] developed a multiple equipment position monitoring method using 3D coordinates. These studies can timely and accurate$l+y$ track construction equipment and obtain their trajectories. However, when construction equipment are close to each other or workers, only treating the construction equipment as a point will lead to the loss of information, which cannot be effectively described its spatial–temporal information.

The enclosing contour of the construction equipment using HBB can provide more information than the aforementioned point-represented construction equipment methods, in addition to the trajectory there are time-varying width and height. Zhu et al. [24] presented an automatic construction equipment detection and tracking method using HBBs for better precision and recall. Kim and Chi [20] adapted a 2D long-term construction equipment tracking method integrated with real-time online learning-based detector and tracker. Kim and Chi [21] also conducted researches on excavator and truck tracking method based on cross-camera matching techniques. Chen et al. [19] proposed a detection and tracking method for construction equipment to recognize their activities. Xiao and Kang [22] developed a construction equipment tacker using deep learning detector integrated technique. They [23] also proposed a robust night-time construction equipment tracker using deep learning illumination enhancement. These HBB-based tracking methods can reflect the size changes of the construction equipment. But when the aspect ratio of the construction equipment is much greater than 1 or the spatial distribution is dense, the HBB-based enclosing contour would contain a lot of non-target information. Wang et al. and Bang et al. [1, 25] employed instance segmentation method to extract the pixel-level segmented contours of construction equipment. This is an appropriate way for the construction equipment representation. But robust segmented contour tracking based on deep learning needs complex manual labelling and

temporal contour association, which would be superfluous for the moving pattern recognition and tracking.

Arbitrary-oriented object detection methods

OABB is a rotatable rectangle with one more parameter rotating angle than HBB, which is the basis of arbitrary-oriented object representation. Because the perspective of the overhead-view images can better reflect the moving patterns of targets, the basic five parameters can be extracted from images intuitively and accurately, so OABB is more used to detect the enclosing contour of targets in overhead-view images [26, 27].

In recent years, many researchers have devoted their efforts on five-parameter detection based on OABBs. In overhead-view images, targets are distributed with random orientations, which makes detecting targets in this field challenging. Chen et al. [28] designed a OABB-based detection model consisted of two CNN networks, in which one CNN was for arbitrary-oriented regions with the orientation information and the other was for object recognition with multi-level feature extraction. Ma et al. [29] proposed a two-stage multi-oriented detector based on CNN in optical remote sensing images using for OABB prediction. Guo et al. [26] developed a single-stage orientation-aware construction vehicle precise detection approach using CNN with feature fusion technique.

Research challenges and objectives

As mentioned before, vision-based enclosing contour tracking of construction equipment is an important mean to obtain spatial–temporal information in large construction sites. The current vision-based construction equipment tracking methods needed to be strengthened in two aspects: in addition to the translation and size change information obtained by the point-represented or HBB-represented tracking methods, the rotation information should be included; considering the complex manual labelling and temporal association in the pixel-level segmented contour, the concise tracking methodology balancing the accuracy and complexity should be considered.

The objective of this study is to develop an enclosing contour tracking method of construction equipment to acquire not only moving trajectories but also temporal sizes and rotating angles. OABB instead of HBB was employed to establish the robust and accurate tracking model for construction equipment using UAV surveillance videos.

Methodology

In this section, the three modules of the OABB based tracking method of construction equipment, including enclosing contour detection, enclosing contour update, and tracking ID managing, are described in detail. Firstly, the enclosing contour is parameterized using five variables of OABB, a CNN-based contour detection model with multi-level features is built and the loss function is defined; secondly, the video frames are input to the model to get detected contours, and the motion model of the construction equipment is built to get predicted contours, tracked contours are updated from predicted contours using the detected contours; finally, the intersection over union (IOU) of OABBs is used to add, keep or delete multiple construction equipment IDs to obtain the tracking status of each equipment.

Enclosing contour detection

The CNN-based detection module describes the construction equipment in images by OABB enclosing contours. Figure 1 shows the difference between HBB and OABB. HBB is defined by four parameters: centre point coordinate (x, y), width (w), and height (h), while OABB is defined by five parameters: x, y, w, h and rotating angle (r). Figure 1(b) compares the effects of equipment representations with two kinds of bounding box. The enclosing contour detection model, which aims to generate and regress OABBs, is modified from the CenterNet [30]. The model consists of two parts: backbone and detection head, as shown in Fig. 2.

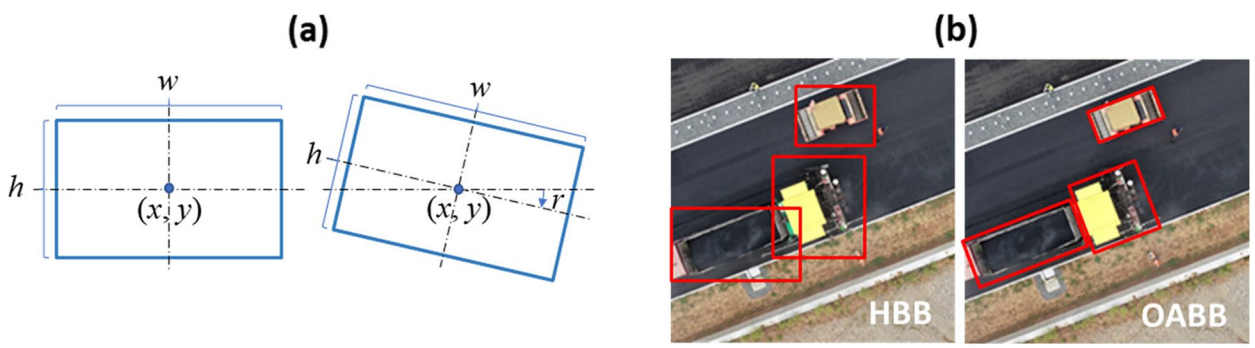

Fig. 1 Difference between HBB and OABB: (**a**) description parameters, (**b**) equipment representations

Backbone provides multi-level features of construction equipment. A modified ResNet-18 base network (mResNet-18) is employed with four residual blocks, each comprising four convolutional layers with two shortcut connections. The residual network has a better fitting ability for extracting more accurate features, and it can also solve the problem of optimisation training when the number of layers increases. Four deconvolution layers are added to recover the spatial information. To speed up the detection efficiency, the output size of the mResNet18 is M / 4 × N / 4 (the size of the input image is M × N).

There are four regression parts in the detection head based on the OABB: centre point regression (x, y), offset regression (off_x, off_y), width and height regression (w, h), and angle regression (r). The four regression parts aim to learn the integers of the centre point coordinates, decimals of the centre point coordinate, width and height, and rotating angle of the OABBs with feature maps processed by $(3 \times 3 \times 64,\ 1 \times 1 \times 2)$, $(3 \times 3 \times 64,\ 1 \times 1 \times 2)$, $(3 \times 3 \times 64,\ 1 \times 1 \times 2)$, and $(3 \times 3 \times 64,\ 1 \times 1 \times 1)$ convolutional kernels, respectively. In the network inference stage, the heat maps from the centre point regression are processed based on 3 × 3 max-pooling, which functions as non-maximum suppression.

To decrease the difficulty of training and increase the efficiency of inference, a Gaussian heat map generated from the ground truth centre point coordinates (x_0, y_0) is employed in this research. w_{xy} and $\hat{w}_{xy}$ are the actual and predicted weights in the Gaussian heat map, respectively. The Gaussian heat map weight at coordinate (x, y) is calculated based on a Gaussian kernel with six parameters: the Gaussian mean (μ_1, μ_2), Gaussian variance (σ_1, σ_1), and window size (r_1, r_2), using Eq. (1), as follows:

$$w_{x,y} = \begin{cases} \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}, x_0 - \frac{r_1}{2} < x < x_0 + \frac{r_1}{2}, y_0 - \frac{r_2}{2} < y < y_0 + \frac{r_2}{2} \\ 0, others \end{cases} \quad (1)$$

$$\mu_1 = x_0, \mu_2 = y_0, \sigma_1 = \lambda w, \sigma_2 = \lambda h, r_1 = 2\sigma_1 + 1, r_2 = 2\sigma_2 + 1$$

The final Gaussian heat map weights at the coordinates (x_g, y_g) are modified based on the rotating angle of the construction equipment as shown in Eq. (2).

$$w_{x_g,y_g} = \begin{cases} w_{x,y}, if \begin{bmatrix} x_g = (x - x_0)\cos ag - (y - y_0)\sin ag + x_0 \\ y_g = (x - x_0)\sin ag + (y - y_0)\cos ag + y_0 \end{bmatrix} \\ 0, others \end{cases} \quad (2)$$

The training loss of the enclosing contour detector (L_{det}, defined by Eq. (3)) is divided into four components, designed based on the detection head: the centre loss (L_c), offset loss (L_o), width and height loss (L_{wh}), and angle loss (L_{ag}). λ_c, λ_o, λ_{wh}, and λ_{ag} are the corresponding weights, respectively. The centre loss employs focal loss for better training convergence, as controlled by Eq. (4), where α and β are adjustment parameters, and N is the number of heat map points, and the other three employ the L1 loss to regress the corresponding parameters.

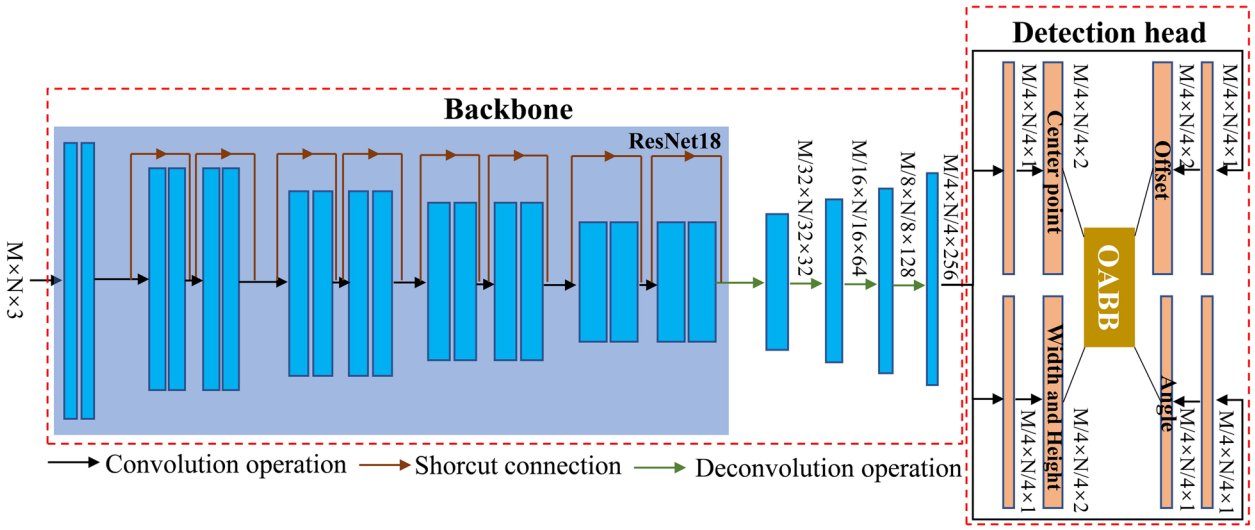

Fig. 2 Detailed architecture of the anchor-free equipment OABB detector

The enclosing contour detection model is pretrained by construction equipment in MOCS proposed by An et al. [31]. For better generalization, the trained network is then fine-tuned by the collected overhead-view construction equipment dataset. The images of this dataset are captured by drone-borne cameras at different heights and angle, containing 600 images and 1570 equipment.

$$L_{\text{det}} = \lambda_c L_c + \lambda_o L_o + \lambda_{wh} L_{wh} + \lambda_{ag} L_{ag} \tag{3}$$

$$L_c = \frac{1}{N} \sum_{xy} \begin{cases} (1-\hat{w}_{xy})^{\alpha} \log(\hat{w}_{xy}), \text{ if } w_{xy} = 1 \\ (1-w_{xy})^{\beta} (\hat{w}_{xy})^{\alpha} \log(1-\hat{w}_{xy}), \text{ otherwise} \end{cases} \tag{4}$$

Enclosing contour update

The detection module could generate high-confidence enclosing contour of construction equipment at each frame without considering the temporal context information, resulting in an inability to match construction equipment between different frames. Inspired by Bewley et al. [32], this module employs a Kalman filter to model the frame-by-frame enclosing contours from detection module in the time domain. The Kalman filter predicts the enclosing contours based on the previous contours, and weights the predicted contours with the detected contours for much more accuracy. The state variables of OABB-based construction equipment motion (translation, size change and rotation) can be described as shown in Eq. (5):

$$\mathbf{x} = \left[c_x, c_y, w, h, r, c_x', c_y', w', h', r'\right]^{\mathrm{T}} \tag{5}$$

where c_x',c_y',w',h' and r' are the first derivatives of the corresponding OABB parameters. Assuming that the construction equipment is moving at a relatively low speed (reasonable for equipment at construction sites), the size and orientation of the equipment will change uniformly over a short time Δt. The state function describing OABB-based construction equipment motion could be expressed as Eq. (6):

$$\hat{\mathbf{x}}_{k|k-1} = \begin{bmatrix} 1&0&0&0&0&\Delta t&0&0&0&0 \\ 0&1&0&0&0&0&\Delta t&0&0&0 \\ 0&0&1&0&0&0&0&\Delta t&0&0 \\ 0&0&0&1&0&0&0&0&\Delta t&0 \\ 0&0&0&0&1&0&0&0&0&\Delta t \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&1 \end{bmatrix} \begin{Bmatrix} c_{x,k-1} \\ c_{y,k-1} \\ w_{k-1} \\ h_{k-1} \\ r_{k-1} \\ c'_{x,k-1} \\ c'_{y,k-1} \\ w'_{k-1} \\ h'_{k-1} \\ r'_{k-1} \end{Bmatrix} = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{w}_{k-1} \tag{6}$$

where $\mathbf{x}_{k-1}$ represents the construction equipment state at the (k-1)$^{\text{th}}$ frame and $\hat{\mathbf{x}}_{k|k-1}$ is calculated state estimation at the k^{th} frame using $\mathbf{x}_{k-1}$ and state function; Δt is the time interval of per frame, and $\mathbf{F}$ is the state transition matrix; $\mathbf{w}_{k-1}$ indicates process noise of the investigated equipment motion model, assumed to be white noise with 0 mean and $\mathbf{Q}_{k-1} = E\left(\mathbf{w}_{k-1}\mathbf{w}_{k-1}^{\mathrm{T}}\right)$ covariance. The covariance estimation of the state variables, described by the state covariance matrix $\mathbf{P}$, can be obtained by linearization of the equipment motion model from Eq. (7):

$$\hat{\mathbf{P}}_{k|k-1} = \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^{\mathrm{T}} + \mathbf{Q}_{k-1} \tag{7}$$

where $\hat{\mathbf{P}}_{k|k-1}$ illustrates the predicted state covariance matrix using optimal estimation $\mathbf{P}_{k-1}$ and the investigated equipment motion model.

In Kalman prediction stage, the predicted contours have certain difference with actual situations. Therefore, at this stage, the contour information of the detected construction equipment would be used as the measured value(s) for the Kalman update. The state transition from the state vector to the measurements is shown in Eq. (8), where $\mathbf{z}_k$ is the measurement of the k^{th} frame, and $\mathbf{H}$ is the measurement matrix. Only the former five parameters can be acquired from the actual detected contours; thus, the size of $\mathbf{H}$ is 5×10.

$$\mathbf{z}_k = \begin{bmatrix} 1&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0 \end{bmatrix} \begin{bmatrix} c_{x,k} \\ c_{y,k} \\ w_k \\ h_k \\ r_k \\ c'_{x,k} \\ c'_{y,k} \\ w'_k \\ h'_k \\ r'_k \end{bmatrix} = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k \tag{8}$$

where $\mathbf{v}_k$ represents the measurement noise, assumed to be white noise with 0 mean and $\mathbf{R}_k = E\left(\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}\right)$ covariance. The Kalman gain ($\boldsymbol{\kappa}$), calculated using Eq. (9), is the core matrix in the Kalman filter, considering both the prediction and the measurements to update

$$\boldsymbol{\kappa}_k = \hat{\mathbf{P}}_{k|k-1}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\hat{\mathbf{P}}_{k|k-1}\mathbf{H}^{\mathrm{T}} + \mathbf{R}_k\right)^{-1} \tag{9}$$

Using the Kalman gain, the state vectors and state variances of the construction equipment from the Kalman prediction can be updated using Eqs. (10) and (11). And the updated OABB information of the construction equipment considering temporal detection information can be set as the final tracked enclosing contour of the k^{th} frame.

$$\mathbf{x}_k = \hat{\mathbf{x}}_{k|k-1} + \kappa_k\left(\mathbf{z}_k - \mathbf{H}\hat{\mathbf{x}}_k\right) \tag{10}$$

$$\mathbf{P}_k = (\mathbf{I} - \kappa_k \mathbf{H})\hat{\mathbf{P}}_{k|k-1} \tag{11}$$

Tracking ID managing

The allocation of construction equipment IDs is a core issue in multiple construction equipment tracking. Most HBB-based tracking methods lead to the overlapping of boxes for multiple objects, resulting in a high complexity in the data associations between frames. For the OABB represented construction equipment, there is hardly no overlap between the OABBs. Therefore, this research employs the IOU of the OABB as the indicator for the ID managing part (calculated by Eq. (12)).

$$I(U(a,b) = \frac{OABB_a \bigcap oABB_b}{OABB_a \bigcup AAB_b} \tag{12}$$

The ID allocation of construction equipment can be divided into three states: add, keep, and delete. The result of the detected contours and that of the predicted contours are used to calculate the IOU. When the ratio is greater than the pre-setting threshold (IOU_t), the situation is denoted as 'matched'; otherwise, it is denoted as 'unmatched'. When there is an unmatched detected contour and the situation lasts for three consecutive frames, a new equipment ID should be added. When there is an unmatched predicted contour and the situation lasts for three consecutive frames, the corresponding equipment ID should be deleted. The matched detected OABB is used as the measurement for participating in the Kalman update to generate the final tracked contour, and the corresponding equipment ID is maintained.

Evaluation and implementation details

Dataset description

This dataset contains five video clips in various construction environments, captured by cameras mounted on UAVs. All videos were captured in 1080×1080 pixels and filmed at 30 frames per second (FPS) at different heights and view angles. The dataset includes single and multiple equipment, static and moving equipment, hovering and fast-moving cameras, with a total length of 4325 frames, 18 equipment, and 8174 contours, typical frames of evaluation videos are shown in Fig. 3. A detailed description is provided in Table 1. For convenience, annotation was performed every 10 frames. The labelling format is as follows: frame number, equipment ID, centre point coordinates, width and height, angle, and category (confidence score).

Evaluation metrics

The multiple object tracking (MOT) challenge [33] is a multiple object tracking benchmark, and is widely used to evaluate tracker performance. The evaluation metrics employed in this research are modified from the MOT challenge.

Multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP) are core evaluation indexes used to jointly measure a tracker's ability to continuously track objects (i.e. accurately determining the number of objects in consecutive frames, and accurately delineating their positions, so as to achieve uninterrupted continuous tracking). MOTA mainly considers the accumulation of object-matching errors in tracking, and mainly includes FP, FN, and IDs (described as Eq. (13)).

$$MOTA = 1 - \frac{\sum(FN + FP + IDs)}{\sum GT} \in (-\infty, 1) \tag{13}$$

FP and FN represent the wrongly tracked equipment and unmatched ground truth equipment in the unmatched status, respectively. IDs denotes the number of ID switches assigned to ground truth equipment, and GT is the total number of ground truth equipment. MOTA measures the performance of trackers in detecting objects and tracking, and is not affected by the detector performance. MOTP reflects the accuracy of determining the object position and size, and is highly

Fig. 3 Example frames of evaluation videos

Table 1 Description of evaluation dataset for overhead-view construction equipment tracking

Name	FPS	Resolution	Length/frame	Equipment	Contours	Description
CVT-01	30	1080 × 1080	1250	1	1250	excavator in operation
CVT-02	30	1080 × 1080	325	2	350	static vehicles from low height
CVT-03	30	1080 × 1080	600	4	1739	road paving equipment
CVT-04	30	1080 × 1080	650	8	1679	static vehicles from high height
CVT-05	30	1080 × 1080	1500	3	2856	equipment in cooperative operation
Total	/	/	4325	18	8174	/

affected by detector performance. The MOTP is calculated using Eq. (14).

$$MOTP = \frac{\sum_{b,a} IOU(a,b)}{\sum_a c_a} \in (0,1) \tag{14}$$

where a is the frame number, b is the equipment number, c_a is the number of trackers in the matched status, and IOU(a,b) is the IOU value of the matched equipment OABBs.

AR represents the mean square error of tracking rotating angles in degrees. MT represents the number of trajectories matching the ground truth successfully in over 80% of the total frames, respectively. RC and PR are the recall and precision, and represent the ratio of TP OABBs to ground truth OABBs and ratio of TP OABBs to all detected OABBs, respectively. Hz is the processing speed of the algorithms, including the detector in this research; which is different from that used in the MOT challenge.

Implementation details

In the enclosing contour detection module, the excavator, truck, loader, roller and concrete mixer truck categories from the MOCS dataset [31] were selected for pretraining with 1000 epochs. The proposed dataset was processed using augmentation techniques, and then was re-trained or fine-tuned using the weights from pretraining. The total re-training epoch was 350, with an initial learning rate of 1.25×10^{-4}, and a 0.1-fold decay was performed at epochs 200 and 300. The loss weights in Eq. (3), i.e. λ_c, λ_o, λ_{wh}, and λ_{ag} were set to 1.0, 1.0, 0.5, and 1.0, respectively. An Adam optimiser was employed in this training with default hyperparameters to achieve better convergence.

In the enclosing contour update module, as shown in Eq. (15), the state covariance matrix $\mathbf{P}_0$ was initiated, and the measurement covariance matrix $\mathbf{R}_k$ was set as the identity matrix. To find the proper parameter of the process covariance matrix $\mathbf{Q}_k$, λ was used to represent the relationship between $\mathbf{R}_k$ and $\mathbf{Q}_k$, and is set as 5.0. IOU_t was set as 0.8.

$$\mathbf{P}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 \end{bmatrix}, \mathbf{R}_k = \mathbf{I}, \mathbf{Q}_k = \lambda \mathbf{I} \tag{15}$$

In the experiments, a modified tracking method from SORT (mSORT) [32] was chosen as the baseline method to compare with the proposed method in this study to test the tracking results of evaluation videos. Because SORT method is one of the state-of-the-art methods in the field of multiple object tracking, characterized by a flexible framework and fast tracking speed. In addition, the mSORT used for comparison with the proposed method employed the same detector backbone, based on HBB generation and regression to detect construction equipment, and was trained using the same dataset. It also used Kalman filtering for HBB prediction of construction equipment and used more complex linear assignment and IOU of HBBs for ID management.

The hardware platform employed mainly includes an Intel Xeon(R) E5-2620 v4 CPU, a Nvidia GTX 1080Ti GPU, and 32 GB of memory.

Results and discussions

Tracking results

The experimental results using the proposed method and the baseline method are shown in Table 2. To better compare the differences between the two methods, Fig. 4 shows the tracking results of five video example frames, where the solid line box represents results from the proposed method and the dashed line box from mSORT. The tracking performance of the five video clips from the evaluation dataset was averaged. The proposed method achieved the recall of 99.381%, precision of 98.165%, MOTA of 97.523%, MOTP of 83.243%. Meanwhile, MT = 18 indicates that the proposed method successfully tracked all 18 trajectories of construction

Table 2 Quantitative evaluation tracking results for the evaluation dataset

Name	Tracking method	RC	PR	MOTA	MOTP	AR	MT	Hz
CVT-01	Proposed method	100	100	100	88.025	2.564	1	33
	mSORT	100	80.985	100	63.224	/	1	33
CVT-02	Proposed method	100	100	100	81.804	2.889	2	32
	mSORT	100	74.267	100	54.398	/	2	33
CVT-03	Proposed method	100	95.714	95.522	84.790	2.029	4	29
	mSORT	100	78.932	93.745	58.753	/	4	30
CVT-04	Proposed method	98.485	97.015	95.455	76.590	4.374	8	28
	mSORT	96.229	76.468	92.698	57.441	/	8	39
CVT-05	Proposed method	99.123	99.123	98.246	84.366	2.322	3	30
	mSORT	97.554	73.815	96.256	56.852	/	3	30
Overall	Proposed method	99.381	98.165	97.523	83.243	2.657	18	29
	mSORT	98.754	72.778	95.763	58.694	/	18	30

equipment. From the tracking results, it can be seen that the proposed method can accurately and robustly track construction equipment from the overhead-view videos. Specifically, the proposed method improves 25.387% over mSORT on precision and 24.549% on MOTP. It is worth noting that the proposed method achieves 97.523% MOTA, which proves high robustness. The MOTP metric can also be improved by improving the backbone with higher feature extraction efficiency and increasing the amount of training data. The overall AR achieved an averaged 2.657 degrees, which validates the effectiveness of the rotation tracking. There is no significant difference between the tracking speed of the proposed method and mSORT, both up to about 30 frames per second, which can be called real-time processing algorithms. If the speed of the algorithm needs to be further increased, it can be done by improving the hardware capability or by using techniques such as parallel coding.

In the evaluation results, CVT-01 contains only one moving construction equipment, and the proposed method achieved 88.025% of MOTP, which improved 24.801% comparing to mSORT. That proves the effectiveness of the proposed OABB for single equipment representation. The two parked construction equipment filmed with a fast-rotating camera are continuously assigned two IDs in CVT-02, with a MOTP of 81.804%. The proposed method improved 27.406% of MOTP than mSORT. CVT-03 contains dense multiple construction equipment and has a construction equipment moving out of view and another equipment moving into view, and the proposed method successfully deleted the ID of the former when it disappeared, and allocated a new ID for the latter with a MOTP of 84.790%. There are eight successive different construction equipment entering in CVT-04 with a MOTP of 76.59%, and the proposed method correctly handles the complex destruction and creation of equipment IDs with accurate detections. The AR achieved 4.374 degrees, which is the highest among the five videos. That validates the difficulty of small equipment rotation identification. CVT-05 contains two construction equipment in cooperative operation; one of them moves out of view, and then moves into view again. The proposed method achieved 84.366% of MOTP. The equipment was allocated to different IDs, because the proposed method could not re-recognise the same equipment which re-entered the view. In conclusion, the tracking results illustrate that the proposed method can accurately detect construction equipment and stably track different equipment, and has a significant improvement on tracking accuracy comparing to mSORT.

Influences of OABB update parameter

The enclosing contour update is conducted by the fusion of detected OABB and predicted OABB. The measurement covariance matrix R represents the detection noise in the equipment OABB generation and regression, which is validated as a high-confidence detector. Thus, R is set to a small value (the identity matrix in this research). The process covariance matrix Q reflects the process noise of the assumed dynamic motion model, and is abstracted from the complex actual situation. λ controls the ratio of Q to R, and Table 3 shows the quantitative evaluation results for different λs.

Table 3 indicates that when λ is greater than or equal to 5.0, that is, the measurement error is relatively small, there is an increase in the *MOTA*, but there are no evident changes in the other indicators. Therefore, in this study, λ is set to 5.0. This experiment also proves that the proposed tracking method is robust to the assumptions of the construction equipment motion model.

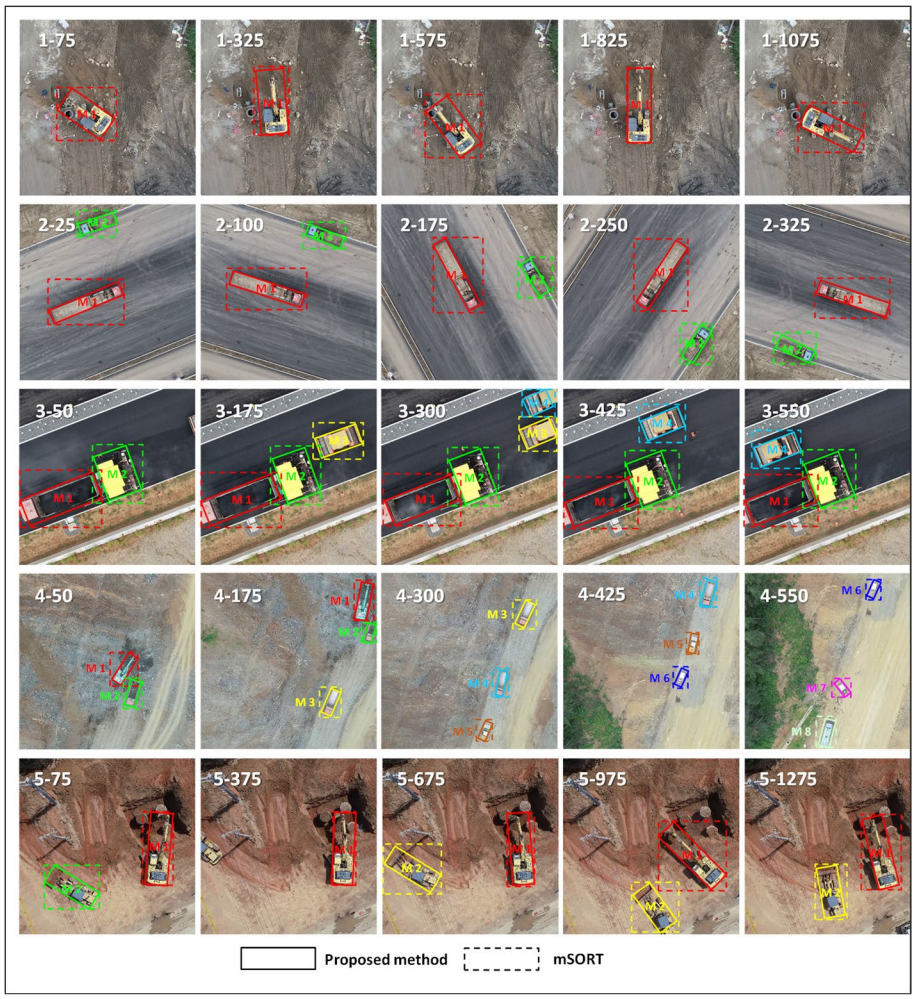

Fig. 4 Tracking results comparison between the proposed method and mSORT

Conclusions and future works

This study proposes a fully automated vision-based enclosing contour tracking method for construction equipment of highway construction sites to obtain the spatial–temporal information of equipment motion. The conclusions could be drawn as follows:

(1) The proposed method integrated OABB to CNN enclosing contour detection of construction equipment; presented a ten-parameter motion model of construction equipment for enclosing contour prediction and updating using Kalman filtering; and finally employed IOU metric instead of complex data association process for ID management of multiple construction equipment.

(2) The proposed method was tested using five evaluation videos, obtaining 2.657 degrees in angle error, 97.523% of MOTA and 83.269% of MOTP, a satisfactory level in multiple object tracking field. And the proposed method could track all 18 trajectories of construction equipment. The experimental results show the advantage

Table 3 Quantitative evaluation tracking results with different λs

λ	RC	PR	MOTA	MOTP	AR	MT	Hz
1.0	97.523	98.438	95.975	83.448	2.657	17	29
5.0	99.381	98.165	97.523	83.269	2.657	18	29
10.0	99.381	98.165	97.523	83.243	2.655	18	29
20.0	99.381	98.165	97.523	83.211	2.655	18	29

of arbitrary-oriented object tracking compared to the widely-used mSORT method.

In this study, the proposed method is suitable for accurate tracking of construction equipment within the field of view. The limitation of this paper is that when the tracked construction equipment gradually moves out of the field of view and then enters the field of view again, the proposed method will renumber the equipment as a new construction equipment, that is, the proposed method does not have the ability to re-identify the equipment. The future work will focus on improving the re-identification capability to track construction equipment in re-entering view. Another future direction is to lightweight the contour detection network which is expected to be deployed on mobile devices.

Abbreviations

OABB	Orientation-aware bounding box
UAV	Unmanned aerial vehicle
HBB	Horizontal bounding box
ID	Identification
CNN	Convolutional neural network
mResNet-18	The modified ResNet-18 base network proposed in this paper
IOU	Intersection over union
MOT	Multiple object tracking
MOTA	Multiple object tracking accuracy
MOTP	Multiple object tracking precision
FP	False Positive
FN	False Negative
TP	True Positive
GT	Ground Truth
AR	The mean square error of tracking rotating angles in degrees
MT	The number of trajectories matching the ground truth successfully in over 80% of the total frames
RC	Recall
PR	Precision
MOCS	A dataset named Moving objects in construction sites
mSORT	The modified tracking method from SORT employed in this paper

Acknowledgements
The authors appreciate the National Natural Science Foundation of China, Heilongjiang Natural Science Foundation and Fundamental Research Funds for Central Universities for support of this research.

Authors' contributions
Yapeng Guo: conduct literature review, build models and analyse, draft the manuscript; Yang Xu: provide assistance on building models; Zhonglong Li, provide assistance on the dataset; Hui Li, refine the manuscript; Shunlong Li, envision the study. The authors read and approved the final manuscript.

Funding
The financial support for this study was provided by the NSFC [Grant Nos. 51922034 and 52278299], the Heilongjiang Natural Science Foundation for Excellent Young Scholars [Grant No. YQ2019E025] and Fundamental Research Funds for Central Universities (Grant No. FRFCU5710051018).

Availability of data and materials
The data and code are available upon request.

Declarations

Ethics approval and consent to participate
Yes.

Consent for publication
Yes.

Competing interests
No.

Received: 4 December 2022 Revised: 29 December 2022 Accepted: 3 January 2023
Published online: 16 January 2023

References

1. Bang S, Hong Y, Kim H (2021) Proactive proximity monitoring with instance segmentation and unmanned aerial vehicle-acquired video-frame prediction. Comput-Aided Civ Infrastructure Eng 36(6):800–816. https://doi.org/10.1111/mice.12672
2. Brilakis I, Park M-W, Jog G (2011) Automated vision tracking of project related entities. Adv Eng Inform 25(4):713–724. https://doi.org/10.1016/j.aei.2011.01.003
3. Sherafat B, Ahn Changbum R, Akhavian R, Behzadan Amir H, Golparvar-Fard M, Kim H, Lee Y-C, Rashidi A, Azar Ehsan R (2020) Automated methods for activity recognition of construction workers and equipment: state-of-the-art review. J Constr Eng Manag 146(6):03120002. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001843
4. Teizer J (2015) Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. Adv Eng Inform 29(2):225–238. https://doi.org/10.1016/j.aei.2015.03.006
5. Yang J, Park M-W, Vela PA, Golparvar-Fard M (2015) Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future. Adv Eng Inform 29(2):211–224. https://doi.org/10.1016/j.aei.2015.01.011
6. Guo H, Yu Y, Skitmore M (2017) Visualization technology-based construction safety management: a review. Autom Constr 73:135–144. https://doi.org/10.1016/j.autcon.2016.10.004
7. Park M-W, Makhmalbaf A, Brilakis I (2011) Comparative study of vision tracking methods for tracking of construction site resources. Autom Constr 20(7):905–915. https://doi.org/10.1016/j.autcon.2011.03.007
8. Seo J, Han S, Lee S, Kim H (2015) Computer vision techniques for construction safety and health monitoring. Adv Eng Inform 29(2):239–251. https://doi.org/10.1016/j.aei.2015.02.001
9. Xu S, Wang J, Shou W, Ngo T, Sadick A-M, Wang X (2021) Computer Vision techniques in construction: a critical review. Arch Comput Methods Eng 28(5):3383–3397. https://doi.org/10.1007/s11831-020-09504-3
10. Arslan M, Cruz C, Roxin A-M, Ginhac D (2018) Spatio-temporal analysis of trajectories for safer construction sites. Smart Sustainable Built Environ 7(1):80–100. https://doi.org/10.1108/SASBE-10-2017-0047
11. Lu M, Chen W, Shen X, Lam H-C, Liu J (2007) Positioning and tracking construction vehicles in highly dense urban areas and building construction sites. Autom Constr 16(5):647–656. https://doi.org/10.1016/j.autcon.2006.11.001
12. Song J, Haas Carl T, Caldas Carlos H (2006) Tracking the Location of Materials on Construction Job Sites. J Constr Eng Manag 132(9):911–918. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:9(911)
13. Kim D, Liu M, Lee S, Kamat VR (2019) Remote proximity monitoring between mobile construction resources using camera-mounted UAVs. Autom Constr 99:168–182. https://doi.org/10.1016/j.autcon.2018.12.014
14. Tang S, Golparvar-Fard M, Naphade M, Gopalakrishna Murali M (2020) Video-Based Motion Trajectory Forecasting Method for Proactive Construction Safety Monitoring Systems. J Comput Civ Eng 34(6):04020041. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000923
15. Zhao Y, Chen Q, Cao W, Yang J, Gui G (2019) Deep learning for risk detection and trajectory tracking at construction sites. IEEE Access 7:30905–30912. https://doi.org/10.1109/ACCESS.2019.2902658
16. Zhu Z, Ren X, Chen Z (2016) Visual tracking of construction jobsite workforce and equipment with particle filtering. J Comput Civ Eng 30(6):04016023. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000573

17. Jog GM, Brilakis IK, Angelides DC (2011) Testing in harsh conditions: Tracking resources on construction sites with machine vision. Autom Constr 20(4):328–337. https://doi.org/10.1016/j.autcon.2010.11.003
18. Zhu Z, Park M-W, Koch C, Soltani M, Hammad A, Davari K (2016) Predicting movements of onsite workers and mobile equipment for enhancing construction site safety. Autom Constr 68:95–101. https://doi.org/10.1016/j.autcon.2016.04.009
19. Chen C, Zhu Z, Hammad A (2020) Automated excavators activity recognition and productivity analysis from construction site surveillance videos. Autom Constr 110:103045. https://doi.org/10.1016/j.autcon.2019.103045
20. Kim J, Chi S (2017) Adaptive detector and tracker on construction sites using functional integration and online learning. J Comput Civ Eng 31(5):04017026. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000677
21. Kim J, Chi S (2020) Multi-camera vision-based productivity monitoring of earthmoving operations. Autom Constr 112:103121. https://doi.org/10.1016/j.autcon.2020.103121
22. Xiao B, Kang S-C (2021) Vision-Based Method Integrating Deep Learning Detection for Tracking Multiple Construction Machines. J Comput Civ Eng 35(2):04020071. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000957
23. Xiao B, Lin Q, Chen Y (2021) A vision-based method for automatic tracking of construction machines at nighttime based on deep learning illumination enhancement. Autom Constr 127:103721. https://doi.org/10.1016/j.autcon.2021.103721
24. Zhu Z, Ren X, Chen Z (2017) Integrated detection and tracking of workforce and equipment from construction jobsite videos. Autom Constr 81:161–171. https://doi.org/10.1016/j.autcon.2017.05.005
25. Wang Z, Zhang Q, Yang B, Wu T, Lei K, Zhang B, Fang T (2021) Vision-based framework for automatic progress monitoring of precast walls by using surveillance videos during the construction phase. J Comput Civ Eng 35(1):04020056. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000933
26. Guo Y, Xu Y, Li S (2020) Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network. Autom Constr 112:103124. https://doi.org/10.1016/j.autcon.2020.103124
27. Ham Y, Han KK, Lin JJ, Golparvar-Fard M (2016) Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works. Vis Eng 4(1):1. https://doi.org/10.1186/s40327-015-0029-z
28. Chen C, Zhong J, Tan Y (2019) Multiple-oriented and small object detection with convolutional neural networks for aerial image. Remote Sensing 11(18):2176. https://doi.org/10.3390/rs11182176
29. Ma J, Zhou Z, Wang B, Zong H, Wu F (2019) Ship detection in optical satellite images via directional bounding boxes based on ship center and orientation prediction. Remote Sensing 11(18):2173. https://doi.org/10.3390/rs11182173
30. Zhou X, Wang D, Krhenbühl P (2019) Objects as Points. arXiv. https://arxiv.org/abs/1904.07850v2
31. An X, Zhou L, Liu Z, Wang C, Li P, Li Z (2021) Dataset and benchmark for detecting moving objects in construction sites. Autom Constr 122:103482. https://doi.org/10.1016/j.autcon.2020.103482
32. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP), 2016. IEEE, pp 3464–3468 https://doi.org/10.1109/ICIP.2016.7533003
33. Milan A, Leal-Taixé L, Reid I, Roth S, Schindler K (2016) MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:160300831. https://arxiv.org/abs/1603.00831

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com